



Workshops

April 2008

Bayesian Analysis of High Dimensional Data

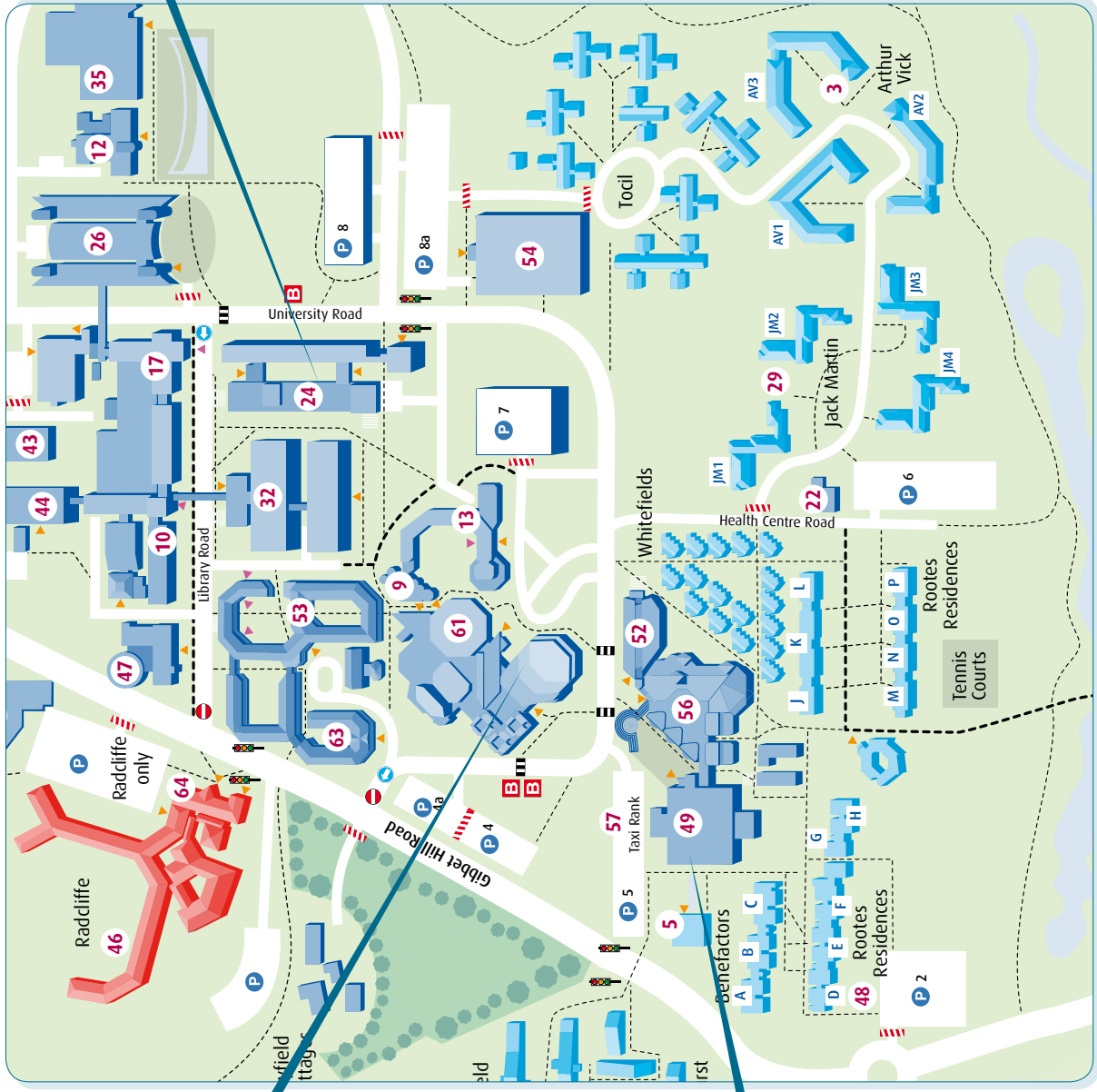
14–16 April 2008

Composite Likelihood Methods

15–17 April 2008

THE UNIVERSITY OF
WARWICK

The central campus of the University of Warwick.



BUILDING KEY

Arthur Wick.....	3
Benefactors.....	5
Chaplaincy.....	9
Computer Science.....	10
Coventry House.....	12
Engineering.....	13
Health Centre.....	17
Humanities Building.....	22
International Manufacturing Centre.....	24
Jack Martin.....	26
Language Centre.....	29
Library.....	24
Mathematics & Statistics.....	35
Physical Sciences.....	43
Physics.....	44
Raddcliffe.....	46
Ramphal Building.....	47
Rootes Residences.....	48
Rootes.....	49
Shops.....	52
Social Studies.....	53
Sports Centre.....	54
Students' Union.....	56
Taxi Rank.....	57
University House.....	60
Warwick Arts Centre.....	61
WBS Social Studies.....	63
WBS Teaching Centre.....	64



24. Humanities



61. Warwick Arts Centre



49. Rootes, Conference Park

TRAVEL INFORMATION

Venue

The workshops take place in the

Mathematics & Statistics Building,
University of Warwick,
Coventry, CV4 7AL,
UK

The workshop venue is building no. **35** on the campus map (upper right corner) and it is also known as “Zeeman Building”.

Check-in for the workshop accommodation is in the Rootes Social Building (building no. **49** on the campus map).

Note:

The University is **not** in the town of Warwick (which is about 8 miles away), and it is **not** the same as Coventry University (which is a different university, located in the centre of Coventry). This is important when telling taxi drivers where you want to be!

Travelling to the University of Warwick

Information on getting to the University of Warwick from Coventry, as well as from other directions locally and further afield, can be found at <http://www2.warwick.ac.uk/about/visiting/directions/> (and via links provided from the navigation bar of that page).

Parking

Residential guests can park in car parks 6, 7, 8 and 15. However, vehicles can be left overnight only in car parks 7 and 15. An exit code for the car parks must be obtained from Rootes Reception at check-in. Car park 15 is located just off the top of the campus map, opposite Building 26.

WELCOME!

Accommodation

Upon arrival, please check in in the Rootes Social Building Reception for instructions (building no. 49 on the campus map). This is also the place to ask about car parking, left luggage, etc.

Workshop registration

Registration for both workshops will take place at the main atrium of the Mathematics & Statistics Building (no. 35 on the map). The registration dates and times are as follows:

- ▷ Bayesian Analysis of High Dimensional Data workshop
Monday 14th April, between 08.30 and 09.20.
- ▷ Composite Likelihood Methods workshop
Tuesday 15th April, between 09.15 and 09.55.

Computing facilities

Computing facilities are available from 09.00–17.00 in room A0.01 of the Mathematics & Statistics Building (located by the main entrance). Alternatively, if you have a WiFi enabled laptop you may access basic internet services¹ by connecting to the “hotspot” wireless network and starting up a web browser. The login details for either A0.01 or wireless are:

	Username	Password
Bayesian workshop	<i>statsws1</i>	<i>roofique</i>
Composite likelihood workshop	<i>statsws2</i>	<i>kiquofot</i>

Meals

During the workshops, lunches and coffee will be provided in the main atrium of the Mathematics & Statistics Building.

For the participants who have registered for dinner, dinners will be served in the Chancellors Suite of the Rootes Social Building (building no. 49) on the 14th and 15th of April and in the Sutherland Suite (same building) on the 13th and 16th of April. The dinner time is 19.00 for both workshops.

For participants who have not registered for dinners, dining options include the *Xanana* bar on the first floor of the Students Union North building (building no. 56), *Bar Fusion* in Rootes, or (until 19.30) the *Café/Bar* in the Warwick Arts Centre. The main entrance of the Union North is next to Costcutter, and the Warwick Arts Centre is just across the road from Rootes and Costcutter.

There is also a bar located on the first floor of Rootes Social Building, which serves draught beers, a good selection of bottled beers, wines, spirits, soft drinks and a variety of teas and coffees. Another possibility is the Varsity pub, which is a 5 minute walk away from the Mathematics & Statistics Building.

¹but note that sending email via SMTP is not allowed

Messages

The telephone number for colleagues or family to leave an urgent message for you during office hours is 02476 574812. For emergency messages outside these times, please call the main University Switchboard on 02476 523523.

Organisers

CRiSM administrator

- ▷ Paula Matthews (paula.matthews@stats.warwick.ac.uk)

Bayesian Analysis of High Dimensional Data

- ▷ David Banks (banks@stat.duke.edu)
- ▷ Jim Griffin (J.E.Griffin-28@kent.ac.uk)
- ▷ Fabio Rigat (f.rigat@warwick.ac.uk)
- ▷ Mark Steel (M.F.Steel@stats.warwick.ac.uk)

Composite Likelihood Methods

Scientific committee

- ▷ Ruggero Bellio (ruggero.bellio@dss.uniud.it)
- ▷ Paul Fearnhead (p.fearnhead@lancaster.ac.uk)
- ▷ David Firth (d.firth@warwick.ac.uk)
- ▷ Nancy Reid (reid@utstat.utoronto.ca)
- ▷ Cristiano Varin (sammy@unive.it)
- ▷ Grace Yun Yi (yyi@uwaterloo.ca)

Local organisers

- ▷ David Firth (d.firth@warwick.ac.uk)
- ▷ Mohand Feddag (M-L.Feddag@warwick.ac.uk)
- ▷ Ioannis Kosmidis (i.kosmidis@warwick.ac.uk)
- ▷ Heather Turner (Heather.Turner@warwick.ac.uk)

If it is necessary to contact the organizers at any time during the workshops you can ring either Fabio Rigat (07947775967) or Ioannis Kosmidis (07849601806).

MEDICAL AND EMERGENCY INFORMATION

Medical assistance

The University Health Centre is open Monday-Friday 09.00-13.00 and 14.00-17.00. Visitors in need of emergency assistance should contact Security on internal extension 22222.

Emergency services and fire procedures

For help in an emergency dial 22222 from any internal telephone and your call will be directed appropriately. Visitors are asked to familiarise themselves with the University's fire procedures which are displayed in each bedroom.

On discovering a fire in a building:

Raise the alarm by breaking the glass in the nearest *Break Glass* point.

On hearing the continuous ringing of fire bells:

Stop what you are doing.

Leave by the nearest Fire Exit.

Walk calmly, do not run.

Do not stop to collect personal belongings.

Make your way to the nearest evacuation point, standing well clear of the building.

Do not re-enter the building until told to do so by the Fire Service or University Security Staff.

Reporting incidents

In the event of an accident or other incident occurring on University premises, please report immediately to Rootes Reception (Ext. 22280) who will then take appropriate action. Please dial 22222 from an internal phone, or 02476 522222 externally to be in direct contact with our 24-hour security staff.

BAYESIAN ANALYSIS OF HIGH DIMENSIONAL DATA

	Event	Time	Details	
	<i>Registration</i>	08.30–09.20	<i>a Poisson process?</i>	
Monday 14 April (MS.04)	Session B.1.1	09.30–10.30	Doug Nychka	keynote
		10.30–11.10	John Haslett	themed
	<i>Coffee</i>	11.10–11.40	<i>in the main atrium, Maths & Stats</i>	
	Session B.1.2	11.40–12.20	Bruno Sansó	themed
		12.20–13.00	Jonathan Rougier	themed
	<i>Lunch</i>	13.00–14.00	<i>in the main atrium, Maths & Stats</i>	
	Session B.1.3	14.00–15.00	Rob Kass	keynote
		15.00–15.40	Jeffrey S. Morris	themed
	<i>Tea</i>	15.40–16.10	<i>in the main atrium, Maths & Stats</i>	
	Session B.1.4	16.10–16.50	Darren Wilkinson	themed
16.50–17.30		Vilda Purutcuoglu	themed	
<i>Workshop Mixer</i>	17.30–19.00	<i>in the Mathematics Common Room</i>		
<i>Dinner</i>	19.00	<i>in the Chancellors suite, Rootes</i>		
Tuesday 15 April (MS.04)	Session B.2.1	09.30–10.30	Carl Rasmussen	keynote
		10.30–11.10	Yee Whye Teh	themed
	<i>Coffee</i>	11.10–11.40	<i>in the main atrium, Maths & Stats</i>	
	Session B.2.2	11.40–12.20	Jurgen van Gael	themed
		12.20–13.00	Mahmoud Zarepour	themed
	<i>Lunch</i>	13.00–14.00	<i>in the main atrium, Maths & Stats</i>	
	Session B.2.3	14.00–14.30	Yves F. Atchade	contributed
		14.30–15.00	Li Chen	contributed
		15.00–15.30	Alexandra Schmidt	contributed
	<i>Tea</i>	15.30–16.00	<i>in the main atrium, Maths & Stats</i>	
<i>Poster Session</i>	16.00–18.30	<i>in the main atrium, Maths & Stats</i>		
<i>Dinner</i>	19.00	<i>in the Chancellors suite, Rootes</i>		
Wednesday 16 April (MS.01)	Session B.3.1	09.30–10.30	Nils Hjort	keynote
		10.30–11.10	Feng Liang	themed
	<i>Coffee</i>	11.10–11.40	<i>in the main atrium, Maths & Stats</i>	
	Session B.3.2	11.40–12.20	Longhai Li	themed
		12.20–13.00	Adrian Dobra	themed
	<i>Lunch</i>	13.00–14.00	<i>in the main atrium, Maths & Stats</i>	
	Session B.3.3	14.00–15.00	David Madigan	keynote
		15.00–15.40	Malay Ghosh	themed
	<i>Tea</i>	15.40–16.10	<i>in the main atrium, Maths & Stats</i>	
	Session B.3.4	16.10–16.50	Michele Guindani	themed
16.50–17.30		Volker Schmid	themed	
<i>Dinner</i>	19.00	<i>in the Sutherland suite, Rootes</i>		

Notes

- ▷ Speakers should gauge their presentations so as to leave enough time for discussion within the time allocated. Ideally, about five to ten minutes at the end of each talk should be reserved for questions.
- ▷ The main atrium of the Mathematics & Statistics Building is the open area as you move from the main entrance to the statistics department (where the mural with vertical lines is).
- ▷ The Monday and Tuesday sessions will be held in **MS.04** (second floor). The Wednesday sessions will be held in **MS.01** (ground floor). The posters will be displayed in the main atrium.
- ▷ The Mathematics Common Room is located on the first floor of the Mathematics & Statistics Building and is above the main atrium.
- ▷ Both dining suites are in the Rootes Social Building (no. 49 on the campus map).

Abstracts

Session B.1.1 (09.30–11.10, Mon 14 April)

Reconstructing past climate using hierarchical models

by DOUG NYCHKA

This talk will involve the topics on Bayesian Hierarchical models, paleoclimate, numerical climate models, and spatial statistics.

Modelling the palaeoclimate

by JOHN HASLETT

The world's climate is entering into a period of change described by the IPCC (Nov, 2007) as potentially “abrupt”. Yet we know pitifully little about such changes. The instrumental record on past climates is limited to about two centuries and shows no abrupt changes. The general circulation models (GCMs) used to explore future climates are essentially equilibrium models: given assumed “forcing scenarios”, we can learn something about future “steady states” but very little about transitions. The palaeoclimate provides a complementary source of information. For example, in Jan 2007, Working Group 1 of the IPCC reported: “During the last glacial period, abrupt regional warmings (probably up to 16°C within decades over Greenland) occurred repeatedly over the North Atlantic region”. Yet inference on the palaeoclimate is indirect and uncertain. This paper discusses some of the successes of Bayesian spacetime inference and several of the challenges; a first attempt was reported in Haslett et al (2006), Bayesian Palaeoclimate Reconstruction - with discussion, JRSS(A). Our focus here is on Europe for the period since the onset of rapid deglaciation towards the end of the last glacial stage, a little less than 15,000 calendar years ago. The presentation will convey the general scientific context and will concentrate on the methodological challenges for Bayesian space-time modelling. What information we have is mostly available via biological and chemical proxies. Examples include: changes in pollen composition as found in sediments, for this reflects changes in vegetation and hence in climate; changes in the composition of oxygen isotopes in Greenland ice, for this reflects past evaporation. Here we focus on pollen, although in principle the methodology is capable of application to many compositional proxies. Such data derive from counts of pollen, from

different types of plant, that have been preserved in sedimentary records, in lakes and in bogs. Inference relies on the “modern analogue” hypothesis: the climate 8,000 years ago in, eg Glendalough in Ireland, is like the modern climate somewhere in the Northern Hemisphere. More abstractly, useful information is contained in models of the relationship between climate and pollen composition in the modern world.

Statistically, the simplest version of the problem may be stated as follows. For each of a number of samples (referred to as sites), n^m modern and n^f fossil, vectors of compositional data $p^m = \{p_j^m; j = 1, \dots, n^m\}$ and $p^f = \{p_j^f; j = 1, \dots, n^f\}$, are available for study; these are often referred to as “pollen assemblages” or “pollen spectra”. For the modern sample, vectors of climate data $c^m = \{c_j^m; j = 1, \dots, n^m\}$ are also available as covariates; the climate values for the fossil sites are missing. Ancient sites have spatial and depth (whence time) coordinates; modern sites have no time coordinates. Inference on depth age relationship is informed by radio-carbon dating. The objective is to estimate the missing values and thus to reconstruct the prehistoric climate. As the dimensionalities are high the challenge is very considerable. The key advance is in joint analysis of the many sources of uncertainty. This permits: (a) borrowing strength across multiple sites, and indeed multiple proxies, by a reliance on a degree of spatio-temporal smoothness in climate change; (b) a modular approach, separately focussed on the climate responses of individual taxa and on radiocarbon dating uncertainties, coherently linked by Monte Carlo methods; and (c) the subsequent sampling of joint space-time climate histories. This latter is of great importance; it directly addresses the need for detailed information on space-time changes, and it does so in the context a careful analysis of the many uncertainties. Specific technical issues include: (a) the circumventing of MCMC in several modules by exploiting the numerical integration within the GMRF approximations of Rue ; (b) the approximation of certain joint inferences by approximations based on marginal inference; (c) the use of a simple, monotone, continuous and piece-wise linear Markov process for making inferences about the calendar of samples; and (d) the modelling of prior temporal smoothness by long tailed random walks based on Normal Inverse Gaussian increments.

Joint work with:

Andrew Parnell

Michael Salter-Townshend

Session B.1.2 (11.40–13.00, Mon 14 April)

A climatology for north atlantic sea surface temperatures

by BRUNO SANSÓ

We consider the problem of fitting a statistical model to historical records of sea surface temperatures collected sparsely in space and time. The records span the whole of the last century and include the Atlantic Ocean north of the Equator. The purpose of the model is to produce an atlas of sea surface temperatures. This consists of climatological mean fields, estimates of historical trends and a spatio-temporal reconstruction of the anomalies, i.e., the transient deviations from the climatological mean. Our model improves upon the current tools used by oceanographers in that we account for all estimation uncertainties, include parameters associated with spatial anisotropy and non-stationarity, transient and long-term trends, and location-dependent seasonal curves. Additionally, since the data set is composed of four types of measurements, our model also includes four different observational variances. The model is based on discrete process convolutions and Markov random fields. Particular attention is given to the problem of handling a massive data set. This is achieved by considering compact support kernels that allow an efficient parallelization of the Markov chain Monte Carlo method used in the estimation of the model parameters. The model is based on a hierarchical structure that is physically sound,

it is easily parallelizable and provides information about the quantities that are relevant to the oceanographers together with uncertainty bounds. The data set is sufficiently large and the area sufficiently complex and extended to serve as a good testbed for global applications.

Bayesian palaeo-calibration of climate models

by JONATHAN ROUGIER

Modern climate models are “semi-empirical”, in the sense that they contain parameters which are not operationally-defined, and must be inferred. This process is rather informal, since the models themselves are enormously expensive to evaluate. However, at its heart is a comparison of model outputs against observations, and the “tuning” of certain parameters to bring these into better agreement. This process has proceeded over several generations of models, so that it is now very difficult to identify a “hold-out” dataset. Hence there is a strong interest in developing new sources of climate data: palaeo-data is one such source. Palaeo-data exists in the form of climate proxies: quantities that are influenced by climate, predominantly pollen counts from lake sediments. Therefore we need an additional modelling step, to map the climate model output to the proxy observations: the proxy model, which is typically cheap to evaluate. We then have a statistical inverse problem: to infer the climate model parameters (and, on the way, the palaeo-climate itself) from the proxy observations. As well as uncertainty in the climate model parameters, this inference must take account of structural errors in the climate model and the proxy model, and observational errors in the proxy data. Neither of the two “standard” approaches will work: the climate model is too expensive to be embedded in a Monte Carlo sampler directly, and too complicated (in the number of its parameters and outputs, and the spatial and temporal interdependence of its outputs) to be statistically-modelled using an emulator. We propose a two step approach which, while not formally coherent, is at least practical. In the first step, we construct a palaeo-version of our climate model, which involves modifying the land-, ice-, and vegetation-maps, and changing the solar forcing. We use this model to construct a Perturbed Physics Ensemble (PPE): a collection of model-evaluations at different sets of parameter values. This PPE can be used to sample palaeo-climates using a Kernel Density Estimator (KDE), and including an extra source of uncertainty to account for the climate model’s structural error. On the basis of this sampling mechanism, we can condition palaeo-climate on the proxy data using the proxy model (which is cheap to evaluate) within a Monte Carlo sampler. In the second step, we treat the reconstructed palaeo-climate as “data”, and use it to reweight the members of the PPE in a second conditioning step. This reweighting translates into an updated distribution on the model-parameters.

Joint work with:

Richardo Lemos

Session B.1.3 (14.00–15.40, Mon 14 April)

Challenges in analyzing neural spike train data

by ROB KASS

One of the most important techniques in learning about the functioning of the brain has involved examining neuronal activity in laboratory animals under varying experimental conditions. Neural information is represented and communicated through series of action potentials, or spike trains, and the central scientific issue in many studies concerns the physiological significance that should be attached to a particular neuron firing pattern in a particular part of the brain. In addition, a major relatively new effort in neurophysiology involves the use of multielectrode recording, in which responses from dozens of neurons are recorded simultaneously. Among other

things, this has made possible the construction of brain-controlled robotic devices, which could benefit people whose movement has been severely impaired.

Scientific questions involving spike trains may be posed in terms of point process intensity functions, and may be answered using Bayesian methods. In my talk I will very briefly outline some of the problems that have been addressed, the progress that's been made, and the challenges to be faced as dimensionality increases.

Bayesian inference for high dimensional functional and image data using functional mixed models

by JEFFREY S. MORRIS

High dimensional, irregular functional data are increasingly encountered in scientific research. For example, MALDI-MS yields proteomics data consisting of one-dimensional spectra with many peaks, array CGH or SNP chip arrays yield one-dimensional functions of copy number information along the genome, 2D gel electrophoresis and LC-MS yield two-dimensional images with spots that correspond to peptides present in the sample, and fMRI yields four-dimensional data consisting of three-dimensional brain images observed over a sequence of time points on a fine grid. In this talk, I will discuss how to identify regions of the functions/images that are related to factors of interest using Bayesian wavelet-based functional mixed models. The flexibility of this framework in modeling nonparametric fixed and random effect functions enables it to model the effects of multiple factors simultaneously, allowing one to perform inference on multiple factors of interest using the same model fit, while borrowing strength between observations in all dimensions. I will demonstrate how to identify regions of the functions that are significantly associated with factors of interest, in a way that takes both statistical and practical significance into account and controls the Bayesian false discovery rate to a pre-specified level. These methods will be applied to a series of functional data sets.

Session B.1.4 (16.10–17.30, Mon 14 April)

High-throughput data for systems biology modelling

by DARREN WILKINSON

Much of computational systems biology is concerned with building dynamic mechanistic models of cellular processes. Post-genomic technologies are able to generate high-throughput data such as time course microarray data that are potentially useful for helping to build and refine such models. This talk will describe some of the challenges associated with Bayesian modelling of high-dimensional time course data from a dynamic mechanistic perspective.

Variational approximation in inference of the kinetic parameters of the MAPK/ERK pathway

by VILDA PURUTCUGLU

The MAPK/ERK pathway is one of the major signal transduction systems which regulates the growth control of all eukaryotes. In our previous study [2], we model this pathway by using a quasi reaction list which consists of 51 proteins and 66 reactions indicating the activation of the system and the degradation of the growth factor's receptor (EGFR). In inference of the stochastic rate constants of associated reactions, we implement the Euler approximation, which is known as the discretized version of a diffusion process, within the Bayesian framework. In estimation via Euler, we update reaction rates and states at discrete time points sequentially. The updates of rates are found via random walk algorithm. But since the diffusion terms of states depend on rate constants in a non-linear way, sampling the candidate states values of the system needs more special MCMC method like Metropolis-within-Gibbs (MWG) technique [3].

In this study our aim is to investigate the application of exact, rather than MWG, sampling by using variational approximation [1]. For this purpose we define an approximate distribution by adding variational parameters in transition kernels and initial state probabilities of the tractable substructure of the true observation matrix used in the estimation. The underlying substructure is generated in such a way that the necessary links between states are decoupled, thereby, Gibbs sampling can be applicable. The lost of information from removing the links is, then, regained by linking the updates of variational parameters at time t . These free parameters are calculated by minimizing the Kullback-Leibler divergence between the true and approximate distribution of states.

Joint work with:

Ernst Wit

References:

- M. I. Jordan (ed.), Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, “An introduction to variational methods for graphical models”, *Learning in Graphical Models*, Cambridge, MIT Press, 1999.
- V. Purutcuoglu and E. Wit, “Bayesian inference for the MAPK/ERK pathway by considering the dependency of the kinetic parameters”, 2007 (under revision).
- D. J. Wilkinson, “Stochastic Modelling for Systems Biology”, Chapman and Hall/CRS, 2006.

Session B.2.1 (09.30–11.10, Tue 15 April)

Data analysis using Gaussian processes

by CARL RASMUSSEN

Gaussian processes (GPs) are well known non-parametric Bayesian models, but surprisingly they are not used extensively in practice. In this mainly methodological talk I'll show three very different types of GP models used commonly in the machine learning community: 1) inferring complex structures in regression via hierarchical learning of covariance functions, 2) non-linear dimensionality reduction using the GP Latent Variable Models (GP-LVM) and 3) GP classification using the Expectation Propagation algorithm. These examples highlight that the fundamental ability of GPs to express and manipulate (Bayesian) distributions over functions make them a powerful, practical foundation for numerous types of applications.

Bayesian language models

by YEE WHYE TEH

Models of sentences from natural languages like English are an important component of many natural language processing technologies, including speech recognition, machine translation, and text retrieval. Due to the complexity and size of natural languages, such language models are necessarily very complex and high dimensional. I will first discuss some recent advances using hierarchical and nonparametric modelling approaches giving state-of-the-art predictive results, then address the problems of adapting language models to specific domains, and the syntactic and semantic representations of words.

Session B.2.2 (11.40–13.00, Tue 15 April)

Beam sampling for infinite hidden Markov models

by JURGEN VAN GAEL

The Infinite Hidden Markov Model (iHMM) [1,2] is an extension of the classical Hidden Markov Model widely used in machine learning and bioinformatics. As a tool to model sequential data, Hidden Markov Models suffer from the need to specify the number of hidden states. Although model selection and model averaging are widely used in this context, the Infinite Hidden Markov Model offers a nonparametric alternative. The core idea of the iHMM is to use Dirichlet Processes to define the distribution of the rows of a Markov Model transition matrix. As such, the number of used states can automatically be adapted during learning; or can be integrated over for prediction. Until now, the Gibbs sampler was the only known inference algorithm for the iHMM. This is unfortunate as the Gibbs sampler is known to be weak for strongly correlated data; which is often the case in sequential or time series data. Moreover, it is surprising that we have powerful inference algorithms for finite HMM's (the forward-backward or Baum-Welch dynamic programming algorithms) but cannot apply these methods for the iHMM. In this work, we propose a method called the Beam Sampler which combines ideas from slice sampling and dynamic programming for inference in the iHMM. We show that the beam sampler has some interesting properties such as: (1) it is less susceptible to strong correlations in the data than the Gibbs sampler, (2) it can handle non-conjugacy in the model more easily than the Gibbs sampler. We also show that the scope of the beam sampler idea goes beyond training the Infinite Hidden Markov Model, but can also be used to efficiently train finite HMM's.

References:

- MJ Beal, Z Ghahramani, CE Rasmussen, The infinite hidden Markov model, *Advances in Neural Information Processing Systems*, 2002.
- YW Teh, MI Jordan, MJ Beal, DM Blei, Hierarchical Dirichlet Processes, *Journal of the American Statistical Association*, 2006.

Series representations for multivariate generalized gamma processes via a scale invariance principle

by MAHMOUD ZAREPOUR

We introduce a scale invariance property for Poisson point processes and use this property to define a series representation for a correlated bivariate gamma process. This approach is quite general and can be used to define other types of multidimensional Levy processes with given marginals. Some important special cases are bivariate G-processes, bivariate variance gamma processes and multivariate Dirichlet processes. Using the scale invariance principle we show how to construct simple approximations to these multivariate processes. To appear in *Statistica Sinica*.

Session B.2.3 (14.00–15.30, Tue 15 April)

An adaptive MCMC sampler for statistical models with intractable likelihood

by YVES F. ATCHADE

We propose a new approach to sample from posterior distributions of statistical models with intractable normalizing constants. The algorithm builds on the extended Wang-Landau algorithm of Atchade and Liu (07) which provides, using a single Monte Carlo run, an efficient estimate of the intractable normalizing constant at every point of the parameter space. We show that the

method is a valid adaptive MCMC method. We illustrate the method with an application to image segmentation.

Joint work with:

Nicolas Lartillot

Christian P. Robert

EAKF-CMAQ: Introduction and evaluation of a data assimilation for CMAQ based on the Ensemble Adjustment Kalman filter technique

by LI CHEN

A new approach is presented for data assimilation using the Ensemble Adjustment Kalman Filter technique for surface measurements of carbon monoxide in a single tracer version of the community air quality model (CMAQ). An implementation of the EAKF known as the Data Assimilation Research Testbed at the National Center for Atmospheric Research was used for developing the model. Three different sets of numerical experiments were performed to test the effectiveness of the procedure and the range of key parameters used in implementing the procedure. The model domain includes much of the northeastern United States. The first two numerical experiments used idealized measurements derived from defined model runs and the last test uses measurements of carbon monoxide from approximately 220 Air Quality System monitoring sites over the northeastern United States, maintained by the Environmental Protection Agency. In each of case the proposed method provided better results than the method without data assimilation.

Spatially hierarchical factor models: building a vulnerability index to dengue fever in Uruguay

by ALEXANDRA SCHMIDT

We introduce a spatially hierarchical factor model, or a vulnerability index, to measure dengue fever in Uruguay. Our proposal combines spatial information among different municipalities across the region (large scale information) with census tracts information at the municipality level (small scale information). *Aedes aegypti*, the main dengue fever transmitting vector, was reintroduced in Uruguay in 1997 with no cases of the disease been registered up to this point in time. It is of great importance to point the regions of the country which are vulnerable to the reintroduction of the disease. It is common to observe, in social studies, social indices built based on sets of indicators observed on census tract level of municipalities across the countries. In our sample the number of census tracts vary significantly, ranging from as low as 16 (Bella Union) up to 1012 (Montevideo) tracts. A simple comparison with a benchmark procedure, one which aggregates observations at the municipal level before building the index, suggests that our vulnerability index is more sensitive to local characteristics and, therefore, more capable of capturing subtle differences across regions of the country. Our factor model entertains point referenced data at the country level and areal data within municipalities. We expect that within a municipality, census tracts which are close together, tend to have similar values of the variables, and behaving on a more independently fashion if the tracts are far apart. On the other hand, in the country level, one expects that index values vary smoothly across the municipalities. More specifically, we combine the information available on p variables measured at n_i , ($i = 1, \dots, n$) census tracts across n municipalities. The municipality size (number of tracts) is taken into account and provide a tool of weighing the contribution of a variable (according to its location) to the overall vulnerability index. Moreover, different from standard procedures, independence across locations is not imposed.

Joint work with:

Hedibert F. Lopes, Esther Salazar, Mariana Gómez, Marcel Achkar

Poster session (16.00–18.30, Tue 15 April)
On the specification of prior distributions for Bayesian variable selection in probit regression with many more variables than observations

by DEMETRIS LAMNISOS

We consider the Bayesian variable selection problem in binary probit regression with many more variables than observations. All the possible subsets of predictor variables determine the model space. Conditional on the model, the ridge prior $N_{p_\gamma}(0, cI_{p_\gamma})$ is frequently placed on the probit regression coefficients, where $N_p(\mu, \Sigma)$ represents a p -dimensional normal distribution with mean μ and covariance matrix Σ and p_γ is the number of included variables in the model. However the hyperparameter c regulates the amount of shrinkage of the regression coefficients and affects the outcome of Bayesian variable selection. Therefore we should be concerned about the specification of the hyperparameter c .

We propose to use the log predictive score to determine its value since c is part of the Bayesian model. The value of c that minimizes the log predictive score is the preferred choice of c . Therefore we have adopted the approach of Gelfand and Dey (1994) who argue that the correct Bayesian approach to determine the model is through predictive distributions. However, in our context, the cross-validation density does not have a closed analytic expression and therefore we show how to employ efficient importance sampling techniques to estimate it. Since cross-validation densities are used to determine c then eventually the resulting variable selection would not only discriminate the sample but also the population into the disease groups. Finally, the importance samplers accuracy in estimating the log predictive score is illustrated using a gene expression data set.

Reference:

Gelfand, A.E. and D.K Dey (1994): “Bayesian Model Choice : Asymptotics and Exact Calculations ,” *Journal of the Royal Statistical Society B*, 56, 501-514.

Bayesian Nonparametric modelling of Grouped Data

by MICHALIS KOLOSSIATIS

Over the last few years, Bayesian nonparametric models have attracted much attention, due in part to the advances in Markov Chain Monte Carlo methods. Often we assume that the data can be modelled by a nonparametric mixture model. The Dirichlet Process (DP) is usually chosen as the prior distribution of the underlying random measure. Several alternative choices have been proposed, including Normalized Random Measures (NRMs), which include the DP and the Normalized Inverse-Gaussian Process (N-IGP). The way to construct NRMs is by normalizing other random measures (for example, the DP can be seen as a normalized Gamma Process and the N-IGP as a normalized Inverse-Gaussian Process). The basic idea is to exploit the infinite divisibility of the unnormalized random measure, in order to construct random probability measures that are identically distributed, but not independent. For example, by using the DP, we can construct a model that is similar to the model proposed by Muller, Quintana and Rosner (2004), but with marginal distributions of the data in both groups that follow a DP. We compare the properties of the two models (in terms of interpretability, theoretical results, computational issues, simulation results etc) and, finally, we consider the generalizations of them in the case of more groups.

High-dimensional data imputation and visualisation: a geochemical case study

by MARTIN SCHROEDER

Missing data are a common problem in many real-world high dimensional data sets and many standard methods for data modelling and data visualisation cannot cope with them. Therefore

a two-stage processing of the data is often necessary. One initially analyses and models the data to understand the missing value process and estimate (or impute) the missing values and subsequently models the completed data set in the visualisation process. The treatment of the missing values impacts the completed data producing unpredictable consequences on the visualisation. Recent developments have led to more principled visualisation methods based on density models which permit the treatment of missing data together with the estimation of the model parameters.

In this work we look at the imputation performance of visualisation methods based on density models such as probabilistic PCA, Kernel PCA, Generative Topographic Mapping and Gaussian Process Latent Variable Models. Our benchmark data are based on geochemical properties of crude oils from the North Sea and Africa. We compare these visualisation-based methods with standard approaches to data imputation including (weighted) mean imputation and iterative multiple regression. We show that the single-stage probabilistic joint imputation-visualisation methods perform better in the presence of missing data than non-probabilistic imputation methods while rendering a two-stage process unnecessary. We also show that measuring the imputation performance of the probabilistic models provides a good measure of the data visualisation quality. In essence we are using the missing data in unlabelled data sets as a cross-validation measure of the goodness of the fit of the density model that underlies the visualisation itself. We speculate on further developments in the treatment of missing data in high dimensional data sets, and their relation to visualisation.

Bayesian inference for the reconstruction of gene regulatory networks with topological constraints

by ANGELA GRASSI

We propose a Bayesian approach for modelling gene regulatory networks with topological constraints, starting from time-course gene-transcription data. We construct a Bayesian hierarchical model in which both the gene interaction matrix and the concentration of regulating proteins are unknown. The identification of the parameters is based on Markov Chain Monte Carlo techniques. A new way to do MCMC inference in high-dimensional problems with complex likelihoods is introduced for inferring the gene interaction matrix.

An adaptive MCMC sampler for statistical models with intractable likelihood

by YVES F. ATCHADE

We propose a new approach to sample from posterior distributions of statistical models with intractable normalizing constants. The algorithm builds on the extended Wang-Landau algorithm of Atchade & Liu (2007) which provides, using a single Monte Carlo run, an efficient estimate of the intractable normalizing constant at every point of the parameter space. We show that the method is a valid adaptive MCMC method. We illustrate the method with an application to image segmentation.

Joint work with:

Nicolas Lartillot

Christian P. Robert

Discovering software bugs with nonparametric Bayesian models

by FINALE DOSHI

We are interested in the problem of identifying bugs in computer programs and additionally patterns of use that cause computer programs to crash. Specifically, we begin with code that has been instrumented to record how many times certain lines are executed during a run through the program. Given successful and failed runs of the program, we wish to identify potential causes of failure (suggested in the Cooperative Bug Isolation project [1]). We assume that the

line counts—which we call software probes—reflect how often certain latent usage patterns are executed during the run. For example, we may expect to see one set of probe counts from a mouse-click, while another pattern of probe counts may correspond to changing a directory.

The Delta-LDA framework of Andrzejewski et. al. [2] first finds usage patterns in successful runs. They next fix these patterns and then try to learn new usage patterns in the failed runs, thus allowing for the fact that failed runs will contain a combination of innocuous patterns as well as the culprit patterns that caused the program to fail. We use a similar approach, but we use an Indian Buffet Process [3] as a prior for the probes that we expect usage patterns to use. This approach allows us to overcome a short-coming in the previous approach, namely, that the number of usage patterns must be prespecified in the Delta-LDA model, and setting that figure a priori is difficult. We develop effective sampling schemes to do inference in this model and demonstrate our approach on both simulated and real data.

Joint work with:

Jurgen Van Gael

References:

- Liblit, B. Cooperative Bug Isolation: Winning Thesis of the 2005 ACM Doctoral Dissertation Competition. Lecture Notes in Computer Science Vol. 4440. Springer 2007.
- Andrzejewski D., Mulhern, A., Liblit, B. Zhu . Statistical Debugging using Latent Topic Models. ECML 2007
- Griffiths, T. and Ghahramani, Z. Infinite Latent Feature Models and the Indian Buffet Process. Gatsby TR 2005-001, 2005.

Bayesian small area estimation with missing data

by VIRGILIO GOMEZ-RUBIO

Statistical bureaus often deal with the problem of estimating different quantities of interest in small areas (Rao, 2003). Examples of these variables include the average income per household and the unemployment rate. Data available usually comes as individual data collected in a survey but area level information (in the form of area totals or averages) is often available from national statistical offices and can be used as auxiliary data.

Bayesian Hierarchical Models (Gelman et al., 1995) provide a unique framework to work with these type of data because different sources of variation can be combined in a suitable way. In addition to covariates, different types of random effects can be included to account for area-to-area variation and spatial autocorrelation.

Given that survey data seldom cover all areas, missing data will appear in most of the areas under study, increasing the number of parameters in the model that need to be estimated. Exploiting area level covariates and spatial correlation is important in order to cope with areas with missing values. In a first approach, we have considered different types of spatial effects that cope with missing data at the area level. In addition to considering models with spatial random effects at the area level we have explored the use of models with spatial random effects at higher administrative levels when there are areas with many neighbours with missing data.

All these methods will be illustrated with an example based on real data on the estimation of the equalised income per household in Sweden at the municipality level in 1992 (Gómez et al., 2008). Small area estimates provided by these methods can be used to identify areas that have particularly low or high values of the target variable. However, the uncertainty associated to the estimates must be taken into account when producing *league tables*.

Joint work with:

N. Best

S. Richardson

P. Clarke

References:

- A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC, Boca Raton, Florida, 1995.
- V. Gómez-Rubio, N. Best, S. Richardson and P. Clarke *Bayesian Statistics for Small Area Estimation*. Working paper available at <http://www.bias-project.org.uk/research.htm>
- J.N.K. Rao. *Small Area Estimation*. Wiley & Sons Inc., 2003.

Bayesian eQTL modelling for association of copy number variation (CNV) and gene expression

by ALEX LEWIN

A recent development in genetics and genomics is the detection of expression QTLs (eQTLs), that is of genes whose mRNA expression level is associated with one or more genetics markers. Such associations are a first step towards understanding complex regulation processes. Data sets for these investigations are generally large, with tens of thousands of transcripts (mRNAs) being measured simultaneously with tens of thousands of markers. The size of these data sets presents challenges to Bayesian methods aiming to coherently analyse markers and transcripts jointly.

Most eQTL experiments have been performed using SNPs (single nucleotide polymorphisms) as markers. These have the advantage that the same SNPs are genotyped for every sample, so the marker data set is regular. In the application we present here, the markers of interest are deletions and duplications of DNA, i.e. Copy Number Variations (CNVs), which are frequently present, for example, in tumour samples. These can be greatly varying in length across different samples, thus creating additional challenges for analysis.

We present a strategy for Bayesian eQTL analysis with CNVs, using a data set consisting of CNVs and expression measured on the same set of 80 cell lines. Our analysis includes standardisation of CNVs across samples, as well as reduction of correlations between markers. The model used for associating CNVs with expression measurements is an extension of the Jia and Xu Bayesian shrinkage model (Jia and Xu, 2007). All transcripts and markers are analysed simultaneously; markers may be associated with multiple transcripts and transcripts may be associated with multiple markers.

In complex analyses such as these, there may in fact be more than one good solution, i.e. the posterior distribution of model parameters may be multimodal. We pay particular attention to the characterisation of different possible solutions, and to good model fit. Our implementation is fast and may be used to analyse thousands of transcripts and markers simultaneously.

Joint work with:

Leonardo Bottolo
Sylvia Richardson

Reference:

- Jia and Xu 2007, *Genetics* **176**, 611-623.

Empirical Bayes - Point process statistical framework for improved inference from single-neuron multiple-trials observations

by GABRIELA CZANNER

Recording single neuron activity from a specific brain region across multiple trials in response to the same stimulus or execution of the same behavioral task is a common neurophysiology protocol. We present a state-space generalized linear model (SS-GLM) to formulate a point process representation of between-trial and within-trial neural spiking dynamics in multiple trials. To estimate the model parameters by approximate Expectation-Maximization algorithm we use a recursive point process filter and fixed-interval smoothing algorithm as analogs of Bayes filter

and smoother. We assess model goodness-of-fit using the time-rescaling theorem, Kolmogorov-Smirnov plot and autocorrelation function; and guide the choice of model order with Akaike information criterion. To answer the physiological questions we use the approximate joint distribution of the maximum likelihood estimates of the model coefficients to compute empirical Bayes estimates to avoid multiple hypothesis testing problem. We illustrate our approach in two applications. In the analysis of hippocampal neural activity recorded from a monkey, we use the model to quantify the neural changes related to learning. In the analysis of primary auditory cortical responses to different levels of electrical stimulation in the guinea pig midbrain, we use the SS-GLM method to analyze auditory threshold detection. Our results demonstrate that the SS-GLM is a more informative tool than commonly used histogram-based and ANOVA methods. Further, our findings have important implications for developing theoretically-sound and practical tools to characterize the dynamics of spiking activity.

Pitman-Yor mixture based reconstruction for emission tomography

by ERIC BARAT

We introduce an emission tomography reconstruction algorithm following a nonparametric Bayesian approach. In contrast with the well known and used Expectation Maximization (EM), the proposed technique does not rely on any bulky (and rather *ad hoc*) space discretization. Namely, we cast the spatial emission density reconstruction problem in the more general Bayesian nonparametric *point inverse problem* framework:

$$F(\cdot) = \int_{\mathcal{X}} \mathcal{P}(\cdot|\mathbf{x}) G(d\mathbf{x}) \quad (1)$$

$$Y_i|F \stackrel{\text{i.i.d.}}{\sim} F, \text{ for } i = 1, \dots, n$$

where \mathcal{X} represents the object space, $\mathbf{Y} = \{Y_1, \dots, Y_n\}$ is the observed F -distributed dataset, $\mathcal{P}(\cdot)$ is a known probability distribution and $G(\cdot)$ is a random probability measure which has to be estimated from observation \mathbf{Y} . In the context of emission tomography, Y_i corresponds to the index of the line of response of the i^{th} observed event. For $l = 1, \dots, L$, $\mathcal{P}(l|\mathbf{x})$ is the projection distribution *i.e.* the probability of recording an event in the l^{th} detection unit given an emission located in \mathbf{x} . Finally, $G(\cdot)$ is the spatial emission distribution (the so-called activity concentration).

We propose to model the random distribution $G(\cdot)$ in (1) using a Pitman-Yor Mixture of Normals prior and a Normal-Inverse Wishart model as base distribution for the Pitman-Yor Process. We introduce a data augmentation scheme where the set of hidden variables are the emission locations in the continuous object space for each observed event. Thanks to the data augmentation, we propose a Markov Chain Monte Carlo algorithm which is able to generate draws from the posterior distribution of the spatial intensity. A difference with EM approach is that the estimated spatial intensity is a continuous function while parameters in EM are given by the mean intensity per pixel. Finally, draws from the intensity posterior distribution allow the estimation of posterior functionals like the posterior mean and variance or point credibility intervals. Results are presented for simulated data based on a 2D brain phantom and compared to ML-EM and Bayesian MAP-EM.

Joint work with:

Thomas Dautremer

Fast information criteria for SVM variable selection

by JOHAN VAN KERCKHOVEN

Using support vector machines for classification problems has the advantage that the curse of dimensionality is circumvented. However, it has been shown that even here a reduction of the number of explicative variables leads to better results. For this purpose, we propose two

information criteria which can be computed directly from the definition of the support vector machine. We assess the predictive performance of the models selected by our new criteria and compare them to a few existing variable selection techniques in a simulation study. Results of this simulation study show that the new criteria are competitive compared to the others in terms of out-of-sample error rate while being much easier to compute. When we repeat this comparison on a few real-world benchmark datasets, we arrive at the same findings. We also provide several adaptations to these SVM criteria. The first adaptation allows the criterion to perform variable selection for other, similar parameter estimation methods, such as kernel logistic regression and AdaBoost for classification, and support vector regression. The other proposed adaptation is a different way of penalising the number of included variables, one which more accurately represents the complexity of the model.

Hierarchical evolutionary stochastic search

by LEONARDO BOTTOLO

Multivariate regression models with many responses has attracted the attention of the statistical community in very recent years. A notable example is the paradigm of eQTL analysis, where thousands of transcripts are regressed versus (hundred of) thousands of markers. In this context the usual problem of multimodality of the posterior distribution, when $p \gg n$, is further exacerbated by the dimension of the response matrix, usually $q \gg n$.

In this work we introduce a new searching algorithm called Hierarchical Evolutionary Stochastic Search (HESS) where the responses are linked in a hierarchical way. To reduce the computational burden, most of the regression parameters are integrated out and a novel sampling strategy based on adapting ideas of Evolutionary Monte Carlo has been designed to efficiently sample from the huge parametric space. Simulated and real data sets are analysed to demonstrate the performance of the proposed algorithm when p and q are both larger than n .

Joint work with:

Sylvia Richardson

Enrico Petretto

Bayesian networks for high dimensional experiments

by DEBORA SLANZI

In designing combinatorial complex experiments we encounter the problem of choosing a set of factors and their interactions to achieve a particular functionality for the system and formulate accurate predictions for unknown possible compositions. Biochemical experiments are generally characterized by a large number of factors with a little prior knowledge on the factorial effects. Conventional experimental design approaches handle with difficulties the problem of the high dimensional search space, and factors are frequently cancelled in the design without a deep understanding of their effects. For this reason main effects analyses are generally conducted ignoring the joint action of the factors, with the consequence of deriving unreliable predictive models. In this study we tackle the problem of identifying main effects and interactions among factors by adopting a class of probabilistic graphical models, namely the class of Bayesian Networks (Cowell et al., 1999; Brogini et al., 2004). The biochemical experiments are conducted according to an evolutionary approach based on a genetic algorithm (Forlin et al., 2008) and the resulting data are analysed by learning the factor multivariate probability distribution described by the Bayesian network.

Joint work with:

Irene Poli

Davide De March

Michele Forlin

References:

- A. Brogini, M. Bolzan, D. Slanzi (2004). Identifying a Bayesian Network for the problem Hospital and Families. The analysis of patient satisfaction with their stay in hospital. In Applied Bayesian Statistical Studies in Biology and Medicine. Eds. M. di Bacco, G. D'Amore, F. Scalfari. Kluwer Academic Publisher, Norwell, MA (USA), Cap.4
- R. G. Cowell, A. P. Dawid, S. L. Lauritzen and D. J. Spiegelhalter (1999) Probabilistic Networks and Expert Systems. Springer-Verlag.
- M. Forlin, I. Poli, D. De March, N. Packard, G. Gazzola and R. Serra (2008) Evolutionary experiments for self-assembling amphiphilic systems. *Chemometrics and Intelligent Laboratory Systems*, 90 (2), 153-160.

The mode oriented stochastic search for nondecomposable graphical models

by ALEX LENKOSKI

We consider the use of Gaussian graphical models in high-dimensional settings, both in the classical context of covariance matrix estimation and, with suitable adjustments, as a means of assessing both regression and clustering models when the number of parameters greatly exceeds the number of observations. A primary concern is the overwhelming size of the model space and we develop a novel technique, the Mode Oriented Stochastic Search (MOSS) algorithm, which quickly finds regions of high posterior probability. MOSS is a neighborhood-oriented search algorithm that builds on the Shotgun Stochastic Search algorithm developed in Jones et al. (2005) and Hans et al. (2007).

The use of conjugate priors gives a precise way of scoring models, though with some computational difficulties, as we include nondecomposable graphs in our model space. We review techniques for obtaining normalizing constants of nondecomposable graphs via Monte Carlo integration originally developed in Atay-Kayis and Massam (2005) and derive the Laplace approximation to reduce computational burden. We then extend the graphical model selection process in a novel manner that can be used to compare competing models in either a regression or clustering setting, even when the number of parameters significantly exceeds the number of observations.

After the implementation of MOSS we use the results of Piccioni (2000) to develop a Block Gibbs Sampler, which helps form a model averaged estimator. Comparing results of a simulation study to those reported by Yuan and Lin (2007) we show that the MOSS algorithm performs better than likelihood-based techniques in estimation over a number of models. By considering simulation studies in regression contexts taken from Nott and Green (2004) and George and McCulloch (1993), we show that our new method for accounting for regression model uncertainty based on graphical models performs better than previously proposed techniques.

After verifying the validity of our approach, we apply the MOSS algorithm to datasets from sociology, macroeconomic growth and gene expression. In each case, classical graphical model search is compared to either our new regression or clustering methodology.

Joint work with:

Hélène Massam

Adrian Dobra

References:

- Atay-Kayis A. and Massam, H. (2005). A Monte Carlo Method for Computing the Marginal Likelihood in Nondecomposable Gaussian Graphical Models, *Biometrika*, **92**, 317-335.
- Hans, C., Dobra, A. and West, M. (2007). Shotgun Stochastic Search for “Large p” regression, *J.A.S.A.*, **102**, 507-516.
- Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C. and West, M. (2005). Experiments in Stochastic Computation for High-Dimensional Graphical Models, *Statistical Science*, **20**, 388-400.

- George, E. I. and McCulloch, R. E. (1993). Variable Selection Via Gibbs Sampling, *J.A.S.A.*, **88**, 881-889.
- Nott, D. J. and Green, P. J. (2004). Bayesian Variable Selection and the Swendsen-Wang Algorithm, *Journal of Computational and Graphical Statistics*, **13**, 141-157.
- Piccioni, M. (2000). Independence Structure of Natural Conjugate Densities to Exponential Families and the Gibbs' Sampler, *Scan. J. of Statistics*, **27**, 111-127.
- Yuan, M. and Lin, Y. (2007). Model Selection and Estimation in the Gaussian Graphical Model, *Biometrika*, **94**, 19-35.

High dimensional missing data in survival analysis

by XIAOHUI ZHAO

We consider the Bayesian analysis of a large survival data set with more than 100 explanatory variables and 2025 patients. The aim of the analysis is to produce improved prognostic indices. Missing values occur in many of the variables and omitting cases with missing values would seriously reduce the number of cases available and might distort our inference.

We need to consider how we model the dependence of survival time on so many covariates and, in particular in this poster, how we construct a missing data model for such a large and diverse set of variables. We compare two approaches, a direct approach in which survival is regressed on the covariates and an indirect approach in which both survival and covariates are regressed on latent variables.

A random effects formulation of high-dimensional Bayesian covariance selection

by JESSICA KASZA

In a microarray experiment, it is expected that there will be correlations between the expression levels of different genes under study. These correlation structures are of great interest from both biological and statistical points of view. From a biological perspective, the correlation structures can lead to an understanding of genetic pathways involving several genes, while the statistical interest lies in the development of statistical methods to identify such structures. However, the data arising from microarray studies is typically very high-dimensional, with an order of magnitude more genes being analysed than there are slides in a typical study. This leads to difficulties in the estimation of the dependence structure of all genes under study. Bayesian graphical models can be used in such a situation, providing a flexible framework in which restricted dependence structures can be considered.

Dobra et al (2004) utilise such models in the analysis of the dependence structure of microarray data, using their technique “High-dimensional Bayesian Covariance Selection, or HdBCS”. While this technique allows for the analysis of independent, identically distributed gene expression levels, often the data available will have a complex mean structure and additional components of variance. For example, we may have gene expression data for genes from several different geographical sites. Our inclusion of site effects in the HdBCS formulation allows such data to be combined, and the dependence structure of the genes estimated using all of the data available for each gene, instead of examining the data from each site individually. This is essential in order to obtain unbiased estimates of the dependence structure. This site effects formulation can be easily extended to include more general random effects, so that any covariates of interest can be included in the analysis of dependence structure.

Reference:

- A. Dobra et al. Sparse Graphical Models for Exploring Gene Expression Data. *Journal of Multivariate Analysis* **90** (2004) 196-212.

Multi-Output model with copula functions

by ALESSANDRO CARTA

The aim of this work is to introduce a new econometric methodology to estimate multi-output production frontiers. Our basic idea is to use a flexible multivariate distribution that allows us to estimate a system of simultaneous equations using the composed error form. This multivariate distribution is represented by the copula function that splits the joint distribution function in twoparts: the marginals and the dependence function (the copula). In this way we can link different stochastic frontiers in a system of equations. We apply Bayesian inference and we develop MCMC samplers to deal we this type of multivariate functions.

Joint work with:

Mark Steel

Session B.3.1 (09.30–11.10, Wed 16 April)**Finding influential regressors in $p \gg n$ models**

by NILS HJORT

I work with methodology for generalised linear models where the number p of covariates is larger than the number n of individuals. Methods more general than e.g. those of ridging emerge when priors are used that correspond to mixtures of lower-dimensional structures. These lead to strategies for finding say the k most promising covariates, for k not exceeding a user-defined threshold number k_0 . The methods are applied to survival data in combination with gene-expression data.

Bayesian learning with overcomplete sets

by FENG LIANG

An important problem in statistics is to retrieve a function or a signal from noisy massive data. In contrast to the orthonormal basis traditionally used in function estimation, overcomplete (or redundant) representations have been advocated due to their flexibility and adaptation. Bayesian methods provide several advantages in learning an overcomplete representation: regularization is specified through priors; inferences on hyperparameters are easily obtained via Markov Chain Monte Carlo; probabilistic outcomes provide a full spectrum to summarize the prediction or estimation. Our recent progress on Bayesian inference with overcomplete wavelet dictionary and reproducing kernel Hilbert space will be presented along with examples.

Session B.3.2 (11.40–13.00, Wed 16 April)**Avoiding bias for feature selection**

by LONGHAI LI

For many classification and regression problems, a large number of features are available for possible use — this is typical of DNA microarray data on gene expression, for example. Often, for computational or other reasons, only a small subset of these features are selected for use in a model, based on some simple measure such as correlation with the response variable. This procedure may introduce an optimistic bias, however, in which the response variable appears to be more predictable than it actually is, because the high correlation of the selected features with the response may be partly or wholly due to chance. We show how this bias can be avoided when using a Bayesian model for the joint distribution of features and response. The crucial

insight is that even if we forget the exact values of the unselected features, we should retain, and condition on, the knowledge that their correlation with the response was too small for them to be selected. In this paper we describe how this idea can be implemented for “naive Bayes” models of binary data. Experiments with simulated data confirm that this method avoids bias due to feature selection. We also apply the naive Bayes model to subsets of data relating gene expression to colon cancer, and find that correcting for bias from feature selection does improve predictive performance.

The mode oriented stochastic search for log-linear models with conjugate priors

by ADRIAN DOBRA

We describe a novel stochastic search algorithm for rapidly identifying regions of high posterior probability in the space of decomposable, graphical and hierarchical log-linear models. Our approach is based on the conjugate priors for log-linear parameters introduced in Massam, Liu and Dobra, 2008. We discuss the computation of Bayes factors through Laplace approximations and the Bayesian Iterate Proportional Fitting algorithm for sampling model parameters. We also present a clustering algorithm for discrete data. We compare our model determination approach with similar results based on multivariate normal priors for log-linear models. The examples concern a six-way, an eight-way and a sparse sixteen-way contingency tables. Extensions of this work involve stochastic algorithms for variable selection in regressions with discrete data. Relevant applications include building classifiers from gene expression, SNP and phenotype data.

Session B.3.3 (14.00–15.40, Wed 16 April)

High-dimensional Bayesian classifiers

by DAVID MADIGAN

Supervised learning applications in text categorization, authorship attribution, hospital profiling, and many other areas frequently involve training data with more predictors than examples. Regularized logistic models often prove useful in such applications and I will present some experimental results. A Bayesian interpretation of regularization offers advantages. In applications with small numbers of training examples, incorporation of external knowledge via informative priors proves highly effective. Sequential learning algorithms also emerge naturally in the Bayesian approach. Finally I will discuss some recent ideas concerning structured supervised learning problems and connections with social network models.

Gene expression-based glioma classification using hierarchical Bayesian vector machines

by MALAY GHOSH

In modern clinical neuro-oncology, the diagnosis and classification of malignant gliomas remains problematic and effective therapies are still elusive. As patient prognosis and therapeutic decisions rely on accurate pathological grading or classification of tumor cells, extensive investigation is going on for accurately identifying the types of glioma cancer. Unfortunately, many malignant gliomas are diagnostically challenging; these non-classic lesions are difficult to classify by histological features, thereby resulting in considerable interobserver variability and limited diagnosis reproducibility. In recent years, there has been a move towards the use of cDNA microarrays for tumor classification. These high-throughput assays provide relative mRNA expression measurements simultaneously for thousands of genes. A key statistical task is to perform classification via different expression patterns. Gene expression profiles may offer more information than classical morphology and may provide a better alternative to the classical tumor diagnosis schemes. The classification becomes more difficult when there are more than two cancer types, as with

glioma.

This talk considers several Bayesian classification methods for the analysis of the glioma cancer with microarray data based on reproducing kernel Hilbert space under the multiclass setup. We consider the multinomial logit likelihood as well as the likelihood related to the multiclass Support Vector Machine (SVM) model. It is shown that our proposed Bayesian classification models with multiple shrinkage parameters can produce more accurate classification scheme for the glioma cancer compared to the several existing classical methods. We have also proposed a Bayesian variable selection scheme for selecting the differentially expressed genes integrated with our model. This integrated approach improves classifier design by yielding simultaneous gene selection.

Session B.3.4 (16.10–17.30, Wed 16 April)

The optimal discovery procedure and Bayesian decision rules

by MICHELE GIUINDANI

We discuss an interpretation of the optimal discovery procedure (ODP, Storey 2006) as an approximate Bayes rule in a nonparametric Bayesian model for multiple comparisons. An improved approximation defines a non-parametric Bayesian version of the ODP statistic (BDP). The definition includes multiple shrinkage in clusters. In a simulation study and a data analysis example we show a (small) improvement in frequentist summaries. The BDP allows easy modifications for dependence of the comparisons and other extensions of the ODP.

A Bayesian hierarchical model for the analysis of a longitudinal dynamic contrast-enhanced MRI cancer study

by VOLKER SCHMID

Imaging in clinical oncology trials provides a wealth of information that contributes to the drug development process, especially in early phase studies. This paper focuses on kinetic modeling in DCE-MRI, inspired by mixed-effects models that are frequently used in the analysis of clinical trials. Instead of summarizing each scanning session as a single kinetic parameter such as median K_{tr} across all voxels in the tumor ROI we propose to analyze all voxel time courses from all scans and across all subjects simultaneously in a single model. The kinetic parameters from the usual non-linear regression model are decomposed into unique components associated with factors from the longitudinal study; e.g., treatment, patient and voxel effects. A Bayesian hierarchical model provides the framework in order to construct a data model, a parameter model, as well as prior distributions. The posterior distribution of the kinetic parameters is estimated using Markov chain Monte Carlo (MCMC) methods. Hypothesis testing at the study level for an overall treatment effect is straightforward and the patient- and voxel-level parameters capture random effects that provide additional information at various levels of resolution to allow a thorough evaluation of the clinical trial. The proposed method is validated with a breast cancer study, where the subjects were imaged before and after two cycles of chemotherapy, demonstrating the clinical potential of this method to longitudinal oncology studies.

COMPOSITE LIKELIHOOD METHODS

(I) invited talk, (C) contributed talk

Tuesday 15 April (MS.05)

Registration
09.15–09.55

in the main atrium, Maths & Stats

Session C.1.1
10.00–11.30

Nancy Reid (I), Some questions about composite likelihood methods
Bruce Lindsay (I), Ratios of composite likelihoods

Coffee

in the main atrium, Maths & Stats

Session C.1.2
12.00–13.10

Harry Joe (I), Efficiency of bivariate composite likelihood
Richard Chandler (C), Adjusting mis-specified likelihood functions

Lunch

in the main atrium, Maths & Stats

Session C.1.3
14.00–15.10

Kung-Yee Liang (I), Composite likelihood: Some biomedical applications
Gunnar Hellmund (C), Estimating functions and composite likelihood for Cox point processes

Tea

in the main atrium, Maths & Stats

Session C.1.4
15.40–16.30

Marta Fiocco (C), Two-stage estimation and composite likelihood in the correlated Poisson-frailty model
Niel Hens (C), On the correlated frailty model for bivariate current status data in infectious disease epidemiology

Posters
16.30–18.30

in the main atrium, Maths & Stats

Dinner, 19.00

in the Chancellors Suite, Rootes Social

Wednesday 16 April (MS.05)

Session C.2.1
09.30–11.00

Paul Fearnhead (I), The use of composite likelihood methods in population genetics
Peter Song (I), Composite likelihood approach to gene network construction

Coffee

in the main atrium, Maths & Stats

Session C.2.2
11.30–12.40

Marc Aerts (I), Pseudo-likelihood methodology for marginally, conditionally, and hierarchically specified models
Zi Jin (C), On the asymptotic properties of the signed composite likelihood ratio statistic

Lunch

in the main atrium, Maths & Stats

Excursion
13.30–18.00

Dinner, 19.00

in the Sutherland Suite, Rootes Social

Thursday 17 April (MS.05)

Session C.3.1
09.30–11.05

Neil Shephard (I), Fitting and testing vast dimensional time-varying covariance models
Christophe Andrieu (C), On-line parameter estimation in general state-space models with pseudo-likelihood and particle filters
Chrysoula Dimitriou-Fakalou (C), Modified Gaussian likelihood estimators for ARMA models on Z^d

Coffee

in the main atrium, Maths & Stats

Session C.3.2
11.35–13.15

Subhash Lele (I), Data cloning: easy maximum composite likelihood estimation for correlated data using Bayesian MCMC methods
Samuel D. Oman (C), An alternative to composite likelihood for analyzing large sets of spatially correlated binary data
Jean-François Plante (C), Adaptive nonparametric likelihood weights

Lunch

in the main atrium, Maths & Stats

Session C.3.3
14.00–15.30

Nils Lid Hjort (I), Likelihood, partial likelihood and composite likelihood for Markov chain models
Cristiano Varin (I), Composite likelihood analysis of mixed autoregressive probit models with application to pain severity diaries

Tea, 15.30–16.00

in the main atrium, Maths & Stats

Notes

- ▷ The main atrium of the Mathematics & Statistics Building is the open area just inside the main entrance (where the mural with vertical lines is).
- ▷ The sessions will be held in **MS.05** (second floor) and the posters will be displayed in the main atrium.
- ▷ Both dining suites are in the Rootes Social Building (no. 49 on the campus map).
- ▷ There will be a break in the workshop programme on Wednesday afternoon for an excursion to National Trust houses and gardens at Packwood and Baddesley Clinton. All workshop participants are welcome to go on the excursion. No charge will be made for this. Travel to the houses will be by coach (= bus); it will take about half an hour to get there. The two houses are about 2 miles apart by road. We will visit Packwood first, then Baddesley Clinton (where there is a tea room, as well as the interesting house and gardens). Participants can travel between the two houses either on the coach or on foot (it's a fairly easy walk, about 3 miles mainly along canal towpaths including the flight of locks at Lapworth; it should take around an hour and a quarter. Those choosing to walk between the houses will obviously have rather less time to see the houses at Packwood and Baddesley Clinton.

Abstracts

Session C.1.1 (10.00–11.30, Tue 15 April)

Some questions about composite likelihood methods

by NANCY REID

Drawing heavily on Cristiano Varin's review paper, I will try to give an overview of current research in composite likelihood methods, and to highlight areas where there seem to be unanswered questions.

Reference:

Varin, C. (2008). On composite marginal likelihoods. *Advances in Statistical Analysis* **92**, 1–28.

Ratios of composite likelihoods

by BRUCE LINDSAY

We consider the construction of composite likelihoods from the point of view of optimal statistical efficiency, using the natural efficiency definition for the estimating functions based on the corresponding composite score functions. It can be shown that under special parameter values, and limitations on the order of the marginal likelihoods used, the composite likelihood corresponding to the optimal score is not a product of likelihoods, but rather a particular ratio of composite likelihoods raised to powers. Such a construction raises two questions: first, can we use the insights from the special parameter values to create a more general efficiency conclusion; second, since the construction is based on optimal scores, is it reasonable to assume we are seeking the maximum of the composite ratio likelihood, so that it could be called maximum composite ratio likelihood?

Session C.1.2 (12.00–13.10, Tue 15 April)

Efficiency of bivariate composite likelihood

by HARRY JOE

Composite likelihood, including pairwise likelihood, procedures can be used for multivariate models with regression/univariate and dependence parameters, when the joint multivariate probability or density is computationally too difficult. We have studied composite likelihood methods for some models for clustered/longitudinal data and times series data. In some cases, the asymptotic efficiency of composite likelihood estimation can be studied via the theory of estimating equations. For clustered data with varying cluster sizes, for several models we study asymptotic relative efficiencies for various weighted pairwise likelihoods, with weight being a function of cluster size. For longitudinal data, we also consider weighted pairwise likelihoods with weights that can depend on lag. Based on our analyses, we can give some recommendations on the choice of weights to achieve better efficiency. We also report on some results for bivariate composite likelihood applied to a stochastic volatility model for financial time series.

Adjusting mis-specified likelihood functions

by RICHARD CHANDLER

This talk starts by considering inference for clustered data, such as those arising in longitudinal studies or space-time problems. The work was motivated by the need for computationally efficient inference procedures in large space-time data sets, when interest lies in marginal time series properties and spatial dependence is considered as a nuisance. In such situations it is convenient to base inference on an 'independence likelihood' which ignores spatial dependence; this can be regarded as a very simple form of composite likelihood. A simple adjustment yields an inference function for which the usual chi-squared asymptotics hold. Without the adjustment, the null distribution for nested model comparisons is that of a weighted sum of chi-squared random variables; the adjustment is therefore computationally appealing. It also has some theoretical advantages. Indeed, for some models, it recovers the true log-likelihood asymptotically without needing to specify the within-cluster dependence structure. The idea is, moreover, equally applicable to any other "working" likelihood function, as well as in general situations where inference is based on optimising some (not necessarily likelihood-based) objective function. The talk will conclude with a review of some open questions and suggestions for further research.

Reference:

Chandler, R.E. and Bate, S. (2007). Inference for clustered data using the independence log-likelihood. *Biometrika* 94, pp. 167-183.

Session C.1.3 (14.00–15.10, Tue 15 April)

Composite likelihood: Some biomedical applications

by KUNG-YEE LIANG

In the last several decades, composite likelihood has drawn a good deal of attention as a tool for statistical inference. This approach is particularly appealing when the full log-likelihood function is difficult to compute and/or far from being normal. For the latter, the adequacy of asymptotic approximation for the maximum likelihood estimate maybe in doubt especially when there are additional nuisance parameters to deal with. Composite likelihood approach has the additional advantage for being more robust in that fewer assumptions, compared to the full likelihood approach, are needed to carry out the inference. In this talk, the points noted above

are illustrated through a series of examples we have encountered in the past. In addition, we will present in details how composite likelihood approach maybe applied to case-control studies with ordinal disease categories. A genetic association study of schizophrenia will be presented for illustration.

Estimating functions and composite likelihood for Cox point processes

by GUNNAR HELLMUND

We suggest two types of very general and applicable estimation techniques for spatial Cox point processes: A method based on moments and a method based on a Bernoulli composite likelihood interpretation.

We discuss the asymptotic properties of the estimators, provide results from simulation studies and apply the techniques to a tropical rainforest dataset.

Until now the most applicable methods for estimation in spatial Cox processes have assumed stationarity, and estimation under weaker assumptions has been a very specialized task using different minimum contrast methods. Through the methods presented we therefore hope for a more wider use of spatial Cox point process models in applications of spatial point processes.

Joint work with:

Rasmus Plenge Waagepetersen

Session C.1.4 (15.40–16.30, Tue 15 April)

Two-stage estimation and composite likelihood in the correlated Poisson-frailty model

by MARTA FIOCCO

We describe the use of composite likelihood and two stage estimation for a Poisson-gamma frailty model. This model is introduced to account for between-subjects correlation occurring in longitudinal count data. Henderson & Shimakura (2003) replaced the joint likelihood contribution of a subject by the sum over all pairs of time points, as this sum is easier to evaluate than the original likelihood. The composite likelihood using all pairs of observations still entails a high-dimensional maximization problem. Estimation in this model is facilitated using a two-stage procedure, where the marginal distributions are used to estimate all parameters except the frailty correlation, which can be cast in a generalized linear model framework, and where a second stage with pairs of observations is used to estimate only the correlation (with the estimates from stage one substituted).

The two stage composite likelihood is studied in simulations and found to be efficient compared to the composite likelihood where only one stage estimation is involved (cf. Henderson & Shimakura, 2003). Finally the suggested method is applied to patient-controlled analgesia dataset where the number of analgesic doses taken by each subject in the successive 12 time intervals following abdominal surgery is reported.

Joint work with:

H. Putter

J. C. van Houwelingen

On the correlated frailty model for bivariate current status data with applications in infectious disease epidemiology

by NIEL HENS

Individual heterogeneity (Coutinho et al., 1999) comprises the differences among individuals' susceptibility to acquire infections, often referred to as "frailties". In its origin, studying individual differences was done in the context of susceptibility to death. In epidemic theory, Coutinho et al. (1999) were the first to systematically treat heterogeneity in the acquisition of infections. Individuals are dissimilar in the way they acquire infections. Some individuals are more susceptible than others and will experience infection earlier. These frailties can be partly explained (e.g. by differences in social contacts), but in most cases constitute an "unexplained residual" component. Gaining insight in the frailty to acquire an infection has an important impact on the design and implementation of control strategies (see e.g. Farrington et al., 2001).

The instantaneous per capita rate at which a susceptible person acquires infection, the so-called force of infection (hazard of infection), has been shown to be age-dependent and can be derived through various techniques based on serological sample data (Anderson, 1982). Because of computational ease, Farrington et al. (2001) used a shared gamma frailty to model bivariate serological data, i.e. bivariate current status data. We will show how this model connects to time-to-event data and correlated frailty models. A first result is the un/identifiability of the correlated/shared frailty model for current status data. Insight is gained in the identifiability result for time to event data in the setting of Yashin et al. (1995) and Giard et al. (2002). Secondly, we will show the effect on the estimated heterogeneity and FOI (marginal) parameters using different frailty distributions in a generalized linear mixed model framework.

References:

- Coutinho, F.; Massad, E.; Lopez, L.; Burattini, M.; Struchiner, C. & Azevedo-Neto, R. Modelling Heterogeneities in Individual Frailties in Epidemic Models. *Mathematical and Computer Modelling*, 1999, 30, 97-115
- Farrington, C.P.; Kanaan, M.N. & Gay, N.J. Estimation of the basic reproduction number for infectious diseases from age-stratified serological survey data (with discussion). *Applied Statistics*, 2001, 50, 251-292
- R. M. Anderson. *Population dynamics of infectious diseases, theory and applications*. Chapman and Hall., London, 1982.
- Yashin, A.; Vaupel, J. & Iachine, I. Correlated Individual Frailty: An Advantageous Approach to Survival Analysis of Bivariate Data. *Mathematical Population Studies*, 1995, 5, 145-159
- Giard, N.; Lichtenstein, P. & Yashin, A. A multistate model for the genetic analysis of the ageing process *Statistics in Medicine*, 2002, 21, 2511-2526

Poster session (16.30–18.00, Tue 15 April)

Composite likelihood inference for space-time covariance function

by MORENO BEVILACQUA

We propose a weighted composite likelihood approach for the space and space-time covariance function estimation problem. The weights are chosen following an optimality criteria based on the Godambe information. The estimation method can be considered as a valid compromise between the computational burden of the maximum likelihood approach and the loss of efficiency of the classical weighted least squares approach using the empirical variogram. An identification criterion based on the weighted composite likelihood is also introduced. The effectiveness of the proposed procedure is illustrated through examples, simulation experiments and by reanalysing a data set on Irish wind speeds (Haslett and Raftery, 1989).

Composite likelihood inference for Bradley-Terry models

by MANUELA CATTELAN

The Bradley-Terry model was developed to analyse experiments involving paired comparisons. It has been successfully applied in many fields including experimental psychology, analysis of sports tournaments and genetics. However, the structured Bradley-Terry model suffers from the limitation of not properly taking into account the correlation in the data. Indeed in situations where sets of objects (or players) are compared in pairs there is a pattern of cross-correlations which is not considered by the Bradley-Terry model. Therefore a more realistic model can be generated by including a correlation structure. Unfortunately, such an enhancement makes the model much more cumbersome and the computation of the likelihood function implies the solution of a high dimensional integral which is usually unfeasible. In these circumstances, pairwise likelihood inference might be a sensible alternative. The presentation will show how to modify the Bradley-Terry model in order to account for correlation in the data and how to overcome the estimation problem by means of pairwise likelihood.

Pairwise likelihood for the longitudinal mixed Rasch model

by MOHAND FEDDAG

This work presents an inferential methodology based on the marginal pairwise likelihood approach for the longitudinal mixed Rasch model. This method belonging to the broad class of composite likelihood provides consistent and asymptotically normal estimators. It involves marginal pair probabilities of the responses which has been derived from a normal scale mixture approximation (Monahan and Stefanski, [1]). With a simulation study, this approach is compared to the maximum marginal likelihood (MML), where the integrals are approximated by Gauss Hermite quadrature. Finally, on the real data from a health-related quality of life study, this methodology is illustrated and compared to those obtained by the use of Conquest (Wu et al., [3]), GLLAMM (Rabe-Hesketh et al., [2]) and WinBugs software.

Joint work with:

S. Bacci

References:

- J. F. Monahan and L. A. Stefanski (1989). Normal scale mixture approximations to $F^*(z)$ and computation for the logistic-normal integral, In Handbook of the Logistic Distribution, N. Balakrishnan (ed.), 529-540. New York: Marcel Dekker.
- S. Rabe-Hesketh, A. Pickles and A. Skrondal (2001). GLLAMM manual (Tech. Rep. 2001/01), London: Department of Biostatistics and Computing, University of London.
- M.L. Wu, R.J. Adams and M. Wilson (2005). Acer Conquest: Generalized Item Response Modelling Software [computer software], Melbourne, Victoria: Australian Council for Educational Research Ltd.

Penalized likelihood for estimation: Beyond bias reduction

by PATRICK HO

In estimation theory, an important aspect concerns determining estimators that achieve optimal performance in a given estimation problem. In the context of full exponential family models, the present work aims to reduce if not minimize asymptotically two measures that take both the bias and the variance of maximum likelihood estimates into account: the mean-squared distance and the expected volume of a confidence set. In Poisson log-linear models and Binomial logistic regression for example, the effect would be to impose schemes of simple iterative adjustments to the data in standard regression software.

Partial likelihood for spatio-temporal point processes

by IRENE KAIMI

We propose a partial likelihood as an alternative to the full likelihood for estimation of the parameters of interest in spatiotemporal point process models. We identify the difference in the straightforward application of the partial likelihood method for finite models as opposed to infinite point process models, which require the evaluation of a spatial integral for each event in the dataset.

Focusing on the more challenging case of the latter, we examine two simulated examples. Firstly, through an inhomogeneous Poisson example for which the maximum likelihood estimators can be obtained, we determine the efficiency loss incurred using our proposed method. The resulting maximum partial likelihood estimators using Monte Carlo methods to approximate the spatial integral are assessed for computational and statistical efficiency. Our second simulated example is a model for the spread of an epidemic disease. For this example, we first recover the model parameters through partial likelihood estimation calculating the integral analytically, and then using an appropriate quadrature method.

Finally, we apply the partial likelihood method to a real data example, concerning nesting birds at the Ebro Delta Natural Park in Spain.

Joint work with:

Peter Diggle

Session C.2.1 (09.30–11.00, Wed 16 April)**The use of composite likelihood methods in population genetics**

by PAUL FEARNHEAD

Composite likelihood methods are proving popular within population genetics, due to the large amount of data available, and the resulting difficulty with calculating the full-likelihood for the data. This talk will review applications of composite likelihood methods, with specific emphasis on methods for estimating fine-scale recombination rates. I will discuss the theory underpinning this application, and what further theoretical results would be of importance in real applications.

Composite likelihood approach to gene network construction

by PETER SONG

Gene network provides an important system for the understanding of many underlying biological mechanisms. Time-course microarray data gives rise to a new platform to reveal how gene-gene interactions behave over time. We propose a hidden Markov model (HMM) for measurements of gene expression, in which gene-gene dependency is characterized by transition probabilities. With regard to the inference for topology of a gene network, the curse of high dimensionality impairs most of traditional inference methods. We develop a new approach based on composite likelihood to overcome this difficulty. Composite likelihood may be regarded as a dimension-reduction version of likelihood inference, and hence it is useful to deal with high-dimensional data. Both EM-algorithm and model selection procedures are re-developed for composite likelihood. Simulation studies and real biological data analysis will be used to demonstrate the proposed model and inference method.

Session C.2.2 (11.30–12.40, Wed 16 April)

Pseudo-likelihood methodology for marginally, conditionally, and hierarchically specified models

by MARC AERTS

Full marginal maximum likelihood can become prohibitive in terms of computation when measurement sequences are of moderate to large length. This is one of the reasons why generalized estimating equations (GEE) have become so popular. One way to view the genesis of GEE is by modifying the score equations to simpler estimating equations, thereby preserving consistency and asymptotic normality, upon using an appropriately corrected variance-covariance matrix. Alternatively, the (log-)likelihood itself can be simplified to a more manageable form. This is, broadly speaking, the idea behind pseudo-likelihood (PL). For example, when a full joint density for a vector of ordinal outcomes is considered, calculating the higher-order probabilities, required to evaluate the score vector and the Hessian matrix, can be prohibitive while, at the same time, interest can be confined to a small number of lower-order moments. The idea is then to replace the single joint density by, for example, all univariate densities, or all pair-wise densities over the set of all possible pairs within a sequence of repeated measures.

We formally introduce pseudo-likelihood and study its main properties. Apart from parameter and precision estimation, an hypothesis testing framework is derived. Through examples, we illustrate the concept for conditionally specified models in the first place, and then move on to marginal and random-effects models. In addition to these general classes, we pay attention to the specific cases of: (1) data of a combined nature, where various outcomes of differing data types are considered jointly; (2) high-dimensional data; and (3) incomplete data.

References:

- Aerts, M., Geys, H., Molenberghs, G., and Ryan, L. (2002). *Topics in Modelling of Clustered Data*. London: Chapman & Hall.
- Molenberghs, G. and Kenward, M.G. (2007). *Missing Data in Clinical Studies*. Chichester: John Wiley & Sons.
- Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. New York: Springer.
- Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer.

On the asymptotic properties of the signed composite likelihood ratio statistic

by ZI JIN

Composite likelihood has received increased attention in handling large dimensional data sets when the computation of the full likelihood is highly complex. In particular, we concentrate on composite marginal likelihood, which arises by considering marginal densities. Under the scenario that consistency and normality of the composite likelihood estimator are satisfied, we summarize different composite test statistics with their asymptotic distributions, and outline higher order asymptotic properties of the signed composite likelihood ratio statistic. Some examples are analyzed in detail, and simulation studies are presented to further explore the efficiency of composite likelihood and asymptotic performance of the signed composite likelihood ratio statistic.

Session C.3.1 (09.30–11.05, Thu 17 April)

Fitting and testing vast dimensional time-varying covariance models

by NEIL SHEPHARD

Building models for high dimensional portfolios is important in risk management and asset allocation. Here we propose a novel way of estimating models of time-varying covariances that

overcome some of the computational problems which have troubled existing methods when applied to hundreds or even thousands of assets. The theory of this new strategy is developed in some detail, allowing formal hypothesis testing to be carried out on these models. Simulations are used to explore the performance of this inference strategy while empirical examples are reported which show the strength of this method. The out of sample hedging performance of various models estimated using this method are compared.

Joint work with:

Robert F Engle

Kevin Sheppard

On-line parameter estimation in general state-space models with pseudo-likelihood and particle filters

by CHRISTOPHE ANDRIEU

State-space models are a very general class of time series capable of modelling dependent observations in a natural and interpretable way. While optimal state estimation can now be routinely performed using SMC (sequential Monte Carlo) methods, on-line static parameter estimation largely remains an unsolved problem. We propose here an original simulation-based method to address this problem in a pseudo-likelihood framework that do not suffer from the so-called “degeneracy” problem of SMC-based methods. We focus in particular on an on-line Expectation-Maximization (EM). It is easy to implement, potentially computationally very cheap and enjoys nice stability properties. We also develop novel simple recursions that allow us to estimate confidence intervals on-line. Computer simulations are carried out to evaluate the performance of the proposed algorithms for finite hidden Markov chains, linear Gaussian and stochastic volatility models. In addition we propose a novel theoretical study of our pseudo-likelihood estimator that allows us to characterise precisely the loss of statistical efficiency inherent to such an approach. We also characterise precisely our estimator when our criterion needs to be approximated for practical purposes. We demonstrate through simulation that our bounds are practically relevant since the predicted behaviour of our estimator is indeed observed in practice.

Joint work with:

A. Doucet

V. B. Tadic

Modified Gaussian likelihood estimators for ARMA models on Z^d

by CHRYSOULA DIMITRIOU-FAKALOU

For observations from an auto-regressive moving-average process on Z^d , where Z is the integer number space and d is a any positive integer number of dimensions, we propose a modification of the Gaussian likelihood, which, if maximized, corrects the edge-effects of the estimators derived. The new estimators are both consistent and asymptotically normal for any dimensionality, and they are efficient, if the original process is Gaussian. We have followed the time domain and we have used finite numbers for the corrections per dimension, which are especially made for the case of the auto-regressive moving-average models of fixed order. Our methods can be applied to both unilateral and bilateral processes.

Session C.3.2 (11.35–13.15, Thu 17 April)
Data cloning: easy maximum composite likelihood estimation for correlated data using Bayesian Markov chain Monte Carlo methods

by SUBHASH LELE

Lele et al. (2007) introduced a new statistical computing method, called data cloning, to calculate maximum likelihood estimates and their standard errors for general hierarchical models. Although the method uses the Bayesian framework and exploits the computational simplicity of the Markov chain Monte Carlo (MCMC) algorithms, it provides valid frequentist inferences such as the maximum likelihood estimates and their standard errors. The inferences are completely invariant to the choice of the prior distributions and therefore avoid the inherent subjectivity of the Bayesian approach. The data cloning method is easily implemented using standard MCMC software. In this paper, we extend the use of data cloning to obtain maximum composite likelihood estimators. The method is illustrated for the mixed effects generalized estimating equations setup where only the first two moments are specified.

An alternative to composite likelihood for analyzing large sets of spatially correlated binary data

by SAMUEL D. OMAN

Suppose the relation between binary responses Y_i and corresponding vectors of explanatory variables, both observed at points in a spatial lattice, is given by a hierarchical generalized linear model in which the Y_i are conditionally independent given a latent Gaussian field of dependent components. If the lattice is moderately large, the need to repeatedly evaluate high-dimensional integrals makes exact maximum-likelihood estimation computationally prohibitive.

We describe here an approximate method using independent-block estimating equations (IBEE), and compare it with pairwise composite likelihood (CL; Heagerty and Lele, 1998), as well as estimation under a working assumption of independence (IEE; Heagerty and Lumley, 2000). All three methods estimate standard errors using the “sandwich estimator” combined with window subsampling (Sherman, 1996). Using a probit link, the IBEE method divides the lattice into disjoint blocks assumed to be independent of one another, thus giving an easily inverted block-diagonal “working covariance matrix”. Covariances within the blocks are modeled using Pearson’s approximation to the true covariances among the Y_i resulting from the latent field, and thus can be evaluated without integration. Moreover, they reflect the binary nature of the responses.

In an application to a set of vegetation data observed at 6,000 points, IEE and CL give essentially the same point estimates and standard errors, while the IBEE approach gives smaller estimated standard errors as well as different (and more easily interpretable) point estimates. In addition, the IBEE computations are substantially faster than those of CL.

We present asymptotic efficiency calculations suggesting the IBEE estimator may be more efficient than pairwise CL in other cases as well.

References:

- Heagerty, P. J. and Lele, S. R. (1998). A composite likelihood approach to binary spatial data. *Journal of the American Statistical Association* **93**, 1099–1110.
- Heagerty, P. J. and Lumley, T. (2000). Window subsampling of estimating functions with application to regression models. *Journal of the American Statistical Association* **95**, 197–211.
- Sherman, M. (1996). Variance estimation for statistics computed from spatial lattice data. *Journal of the Royal Statistical Society Series B* **58**, 509–523.

Adaptive nonparametric likelihood weights

by JEAN-FRANÇOIS PLANTE

Suppose that you must infer about a population, but that data from $m - 1$ similar populations are available. The weighted likelihood uses exponential weights to include all the available information into the inference while discounting the contribution of each datum based on its dissimilarity with the target distribution.

We first present a heuristic justification of the weighted likelihood using the entropy maximization principle. This interpretation which could help the intuitive understanding of other forms of likelihood leads us to suggesting a criterion to determine the likelihood weights from the data.

The proposed MAMSE weights are nonparametric and can be used to determine a weighted mixture of empirical distributions. The uniform convergence of this empirical function is proved and used to show the strong consistency of the maximum weighted likelihood estimate based on the MAMSE weights.

We also introduce the maximum weighted pseudo-likelihood that allows fitting copula models to multivariate data. The MAMSE weights are generalized to that case and yield, once again, consistent estimates.

Some simulations illustrating the behavior of the methods will also be presented.

References:

<http://www.stat.ubc.ca/Research/TechReports/techreports/226.pdf>

<http://www.stat.ubc.ca/~jeff/PhD.thesis.pdf>

Session C.3.3 (14.00–15.30, Thu 17 April)

Likelihood, partial likelihood and composite likelihood for Markov chain models

by NILS LID HJORT

In many spatial and spatial-temporal models, and more generally in models with complex dependencies, it may be too difficult to carry out full maximum likelihood (ML) analysis. Remedies include the use of pseudo-likelihood (PL) and composite-likelihood (CL). The present article studies the ML, the PL and the CL methods for general Markov chain models, partly motivated by the desire to understand the precise behaviour of PL and CL methods in settings where this can be analysed. We present limiting normality results and compare performances in different settings. For Markov chain models, the PL and CL methods can be seen as maximum penalised likelihood methods. We find that the CL strategy is typically preferable to the PL, and that it loses very little to the ML, while sometimes earning in model robustness. It has also appeal and potential as a modelling tool. Our methods are illustrated for consonant-vowel transitions in poetry and for analysis of DNA sequence evolution type models.

Joint work with:

Cristiano Varin

Composite likelihood analysis of mixed autoregressive probit models with application to pain severity diaries.

by CRISTIANO VARIN

Longitudinal data with ordinal outcomes routinely appear in medical applications. An example is the analysis of clinical diaries where patients are asked to score the severity of their symptoms. In this framework, we propose a class of models for ordinal repeated responses with subject-specific random effects and distinguished autocorrelation structures for different groups of patients. Since

likelihood inference for the proposed class of models is computationally infeasible because of high dimensional intractable integrals, a pairwise likelihood analysis is preferred. Here, we discuss some refinements of the methodology to improve on its inferential properties. A reduced bias of the variance components is obtained by one-step jackknife, while p-value estimation of the composite likelihood ratio statistics is performed through parametric bootstrap. The method is applied to a large longitudinal study on the determinants of migraine severity.

Joint work with:

Claudia Czado

PARTICIPANTS

Workshop on Bayesian Analysis of High Dimensional Data

Name	Affiliation	Name	Affiliation
Shola Adeyemi	University of Westminster	Jim Smith	University of Warwick
Matthew Ager	King's College London	Dario Spanó	University of Warwick
Cedric Archambeau	University College London	Mark Steel	University of Warwick
Eric Barat	CEA, France	James Sweeney	Trinity college, Ireland
Sourav Basu	Trinity College Dublin, Ireland	Jared Tanner	Univ. of Edinburgh
Leonardo Bottolo	Imperial College, London	Yee Whye Teh	University College London
Alexis Boukouvalas	Aston University	Michael Titterington	University of Glasgow
Nigel Burroughs	University of Warwick	Hans van ouwelingen	Leiden Univ., Netherlands
Ismael Castillo	Vrije Univ. Amsterdam, Neth.	Johan Van Kerckhoven	Kath. Univ. Leuven, Belgium
Li Chen	University of Bristol	Laura Villanova	University of Padua, Italy
Dan Cornford	Aston University	Brandon Whitcher	GlaxoSmithKline
Maria Costa	University of Warwick	Darren Wilkinson	University of Newcastle
Gabriela Czanner	University of Warwick	Xiaohui Zhao	University of Newcastle
Thaís Fonseca	Warwick University	Huiming Zhu	Hunan University, China
Adrian Dobra	Univ. of Washington, USA		
Finale Doshi	Cambridge University		
Ludger Evers	University of Bristol		
Donald Fraser	Univ. of Toronto, Canada		
Guy Freeman	University of Warwick		
John Fry	University of Warwick		
Malay Ghosh	University of Florida, USA		
Virgilio Gomez-Rubio	Imperial College London		
Angela Grassi	Italian Nat. Research Council		
Jim Griffin	University of Kent		
Michele Guindani	Univ. of New Mexico, USA		
John Haslett	Trinity College Dublin, Ireland		
Nils Hjort	University of Oslo, Norway		
Lasse Holmstrm	University of Oulu, Finland		
Merrilee Hurn	University of Bath		
Chaitanya Joshi	Trinity College Dublin, Ireland		
Miguel Juárez	University of Warwick		
Maria Kalli	University of Kent		
Rob Kass	Carnegie Mellon Univ., USA		
Jessica Kasza	Univ. of Adelaide, Australia		
Michalis Kolossiatis	University of Warwick		
Willem Kruijer	Vrije Univ. Amsterdam, Neth.		
Demetris Lamnisis	Warwick University		
Alex Lewin	Imperial College		
Longhai Li	Univ. of Saskatchewan, Canada		
Feng Liang	Duke University, USA		
Silvia Liverani	University of Warwick		
David Madigan	Columbia University, USA		
Steven Miller	Trinity College Dublin, Ireland		
Jeffrey Morris	MD Anderson, USA		
Edward Morrissey	Warwick University		
Douglas Nychka	NCAR, USA		
Michael Osborne	University of Oxford		
Andrey Pavlov	Queen's University, Canada		
Leto Peel	BAE Systems, ATC		
Vilda Purutcuoglu	Middle East Tech. Univ., Turkey		
Bala Rajaratnam	Stanford University, USA		
Carl Rasmussen	University of Cambridge		
Fabio Rigat	University of Warwick		
Jonathan Rougier	University of Bristol		
Bruno Sansó	UC Santa Cruz, USA		
Volker Schmid	Imperial College London		
Alexandra Schmidt	Univ. Fed. Rio de Janeiro, Brazil		
Martin Schroeder	Aston University		
Debora Slanzi	Univ. Ca' Foscari of Venice, Italy		

PARTICIPANTS

Workshop on Composite Likelihood Methods

* indicates a participant in the Bayesian Analysis of High-Dimensional Data workshop (April 14–16) who will join in the last day (April 17) of this meeting.

Name	Affiliation	Name	Affiliation
Aerts, Marc	Hasselt University	Shephard, Neil	University of Oxford
Anaya Karim	The Open University	Smith*, Jim	University of Warwick
Andrieu, Christophe	University of Bristol	Solis-Trapala, Ivonne	Lancaster University
Bellio, Ruggero	University of Udine	Song, Peter	University of Michigan
Bevilacqua, Moreno	Università di Padova	Sun, Ye	University of Toronto
Castillo*, Ismael	Vrije Universiteit Amsterdam	Taylor, Jeremy	University of Michigan
Cattelan, Manuela	University of Padova	Turner, Heather	University of Warwick
Chandler, Richard	University College London	Varin, Cristiano	Ca' Foscari Univ., Venice
Costa, Maria	University of Warwick	Vidoni, Paolo	University of Udine
Cox, David	Nuffield College, Oxford	Yau, Christopher	University of Oxford
Davison, Anthony	EPFL, Switzerland	Yi, Grace	University of Waterloo
De Angelis, Daniela	MRC Biostat. U., Cambridge	Young, Alastair	Imperial College London
Dimitriou-Fakalou, Chrysoula	University College London		
Fearnhead, Paul	Lancaster University		
Feddag, Mohand	University of Warwick		
Fiocco, Marta	Leiden University		
Firth, David	University of Warwick		
Fraser*, Donald	University of Toronto		
Gao, Xin	York University, Ontario		
Gu, Hong	Dalhousie University		
Hellmund, Gunnar	University of Aarhus		
Hens, Niel	Hasselt University		
Hjort, Nils	University of Oslo		
Ho, Patrick	University of Warwick		
Holmes, Chris	University of Oxford		
van Houwelingen*, Hans	Leiden University		
Jenkins, Paul	University of Oxford		
Jin, Zi	University of Toronto		
Joe, Harry	University of British Columbia		
Kaimi, Irene	Lancaster University		
Kasza*, Jessica	University of Adelaide		
Kosmidis, Ioannis	University of Warwick		
Kruijjer*, Willem	Vrije Universiteit Amsterdam		
Lele, Subhash	University of Alberta		
Liang, Kung-Yee	Johns Hopkins University		
Lindsay, Bruce	Penn State University		
Liu, Jiayi	Lancaster University		
Morrissey*, Edward	University of Warwick		
Nielsen, Jan	Univ. Southern Denmark		
Oman, Samuel	Hebrew Univ. Jerusalem		
Osborne*, Michael	University of Oxford		
Pace, Luigi	University of Udine		
Pavlov, Andrey	Queen's Univ., Ontario		
Pierce, Donald	Oregon Health Sci. Univ.		
Plante, Jean-François	University of Toronto		
Reid, Nancy	University of Toronto		
Salvan, Alessandra	University of Padova		

SPONSORS

Bayesian Analysis of High Dimensional Data



EPSRC

Engineering and Physical Sciences
Research Council



INTERNATIONAL SOCIETY FOR BAYESIAN ANALYSIS

Composite Likelihood Methods

